Current Biology, Vol. 13, 1709–1714, September 30, 2003, ©2003 Elsevier Science Ltd. All rights reserved. DOI 10.1016/j.cub.2003.09.005

'Putting the Face to the Voice': Matching Identity across Modality

Miyuki Kamachi,[1,*] Harold Hill,[1] Karen Lander,[2] and Eric Vatikiotis-Bateson[1,3]

[1]ATR
Human Information Science Laboratories
2-2-2 Hikaridai
Kyoto 619-0288
Japan

[2]Department of Psychology
University of Manchester
Oxford Road
Manchester M13 9PL UK
United Kingdom

[3]Department of Linguistics
University of British Columbia
1866 Main Mall
Vancouver, British Columbia V6T 1Z1
Canada

Summary

Speech perception provides compelling examples of a strong link between auditory and visual modalities [1, 2]. This link originates in the mechanics of speech production, which, in shaping the vocal tract, determine the movement of the face as well as the sound of the voice [3, 4]. In this paper, we present evidence that equivalent information about identity is available cross-modally from both the face and voice. Using a delayed matching to sample task, XAB, we show that people can match the video of an unfamiliar face, X, to an unfamiliar voice, A or B, and vice versa, but only when stimuli are moving and are played forward. The critical role of time-varying information is underlined by the ability to match faces to voices containing only the coarse spatial and temporal information provided by sine wave speech [5]. The effect of varying sentence content across modalities was small, showing that identity-specific information is not closely tied to particular utterances. We conclude that the physical constraints linking faces to voices result in bimodally available dynamic information, not only about what is being said, but also about who is saying it.

Results and Discussion

One of the face's many functions is that it comprises the outer surface of the vocal tract. As such, it is closely linked to the voice, and it is this that allows us to "speech read" (we use the term speech reading rather than lip reading, as it more accurately reflects the fact that it is not just the lips that convey useful information [6, 7]). In this paper, we demonstrate that this link between faces and voices is also sufficient to specify identity across a change in modality.

Both auditory and visual information are known to be important for the perception of spoken language. For example, what is seen can affect what is heard, as shown by the McGurk effect [1], in which fusion of auditory /ba/ and visual /ga/ results in a percept different from both — typically /da/ or /tha/. Another example is the ventriloquist illusion, where the cross-modal illusion of location can enhance selective spatial attention to speech sounds [2]. More generally, having the speaker's face visible [3] improves the perception of speech in noise, even for people with normal hearing, by an amount equivalent to about a 10 dB increase in audibility. These effects are indicative of the close links between the voice and facial movement that result from the mechanics of speech production [4, 8]. The extent to which these links convey identity will clearly be determined by the extent to which they reflect individual differences.

While it is clear that we can identify people from both their faces and their voices independently, the links between the face and the voice raise the possibility that there is also overlapping information. Recent evidence from face recognition research suggests that the movement of faces is useful for recognition [9–11], and, as much of this movement is associated with speech, the critical cues may be available bimodally. In the auditory modality, these dynamic variations could correspond to prosodic differences between speakers. Phonetic properties of speech are known to be important to voice recognition in addition to purely acoustic cues based on voice quality [12]. The experiments reported here were designed as tests for the existence of bimodal information of this kind and to specify its critical properties.

In all of the experiments reported, we made use of delayed matching to a sample task, XAB, with a change in sensory modality from auditory to visual or vice versa between the first and second phases. This design is illustrated in Figure 1. A face (or voice), X, was presented in the first phase, and then the observer was presented with two voices (or faces) in the second phase. The task was to choose which of the stimuli in the second phase corresponded to that presented in the first phase.

All the faces and voices used were unfamiliar to the observers to ensure that faces and voices could not just be recognized independently in each modality, thus obviating the need for cross-modally available information. It would also be possible to do the task if other information available from both the face and the voice, for example, gender, ethnicity, or age, was sufficient to constrain the match. To avoid this possibility, all pairs of speakers, A and B, were of the same ethnicity, sex, and age group. We also ensured that what was said did not determine the correct match, as information about speech content may have been available from speechread cues as well as from the voice. All of these possibilities were avoided in the experiments reported here: as Figure 1 shows, A and B were always of different people of the same sex saying the same sentence, while X was of either speaker A or B saying a different sentence (texts for the sentences used are given in the Experimental

*Correspondence: miyuki@atr.co.jp

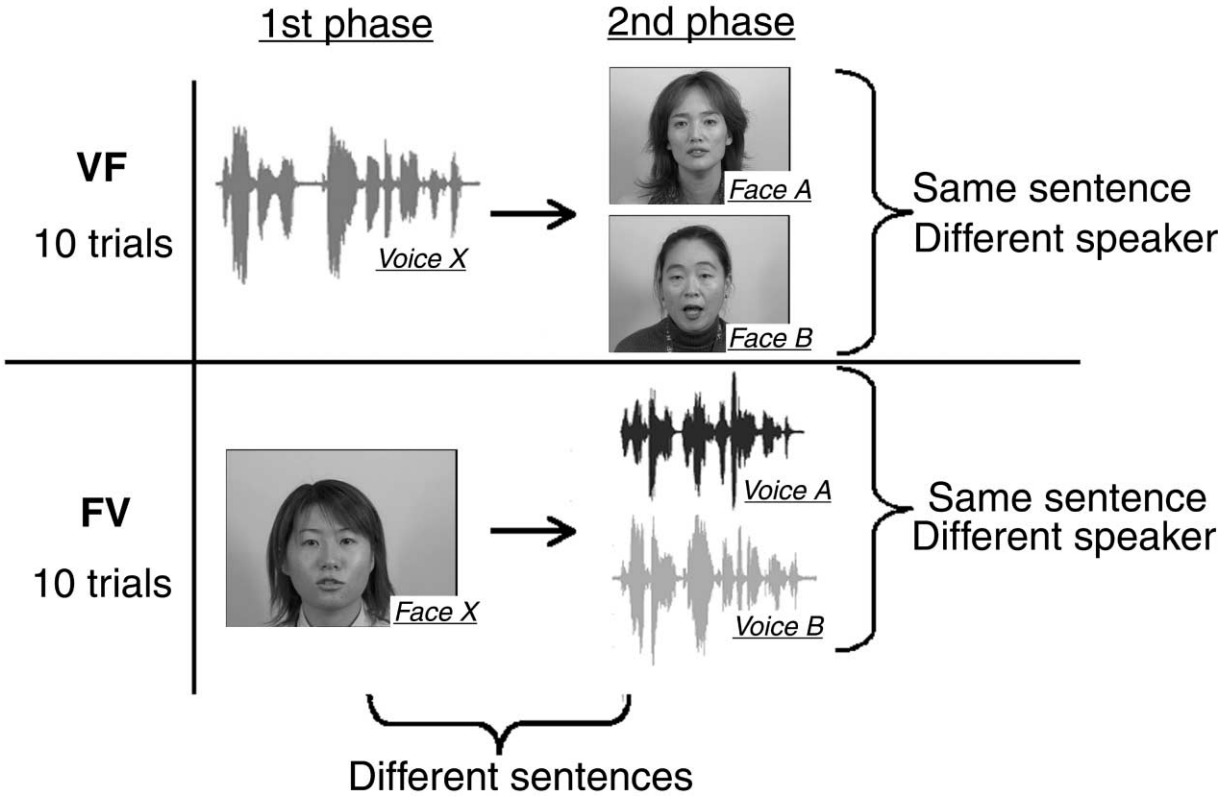

Figure 1. The Fundamental Scheme for Experiment 1

In the VF condition, a voice, X, was learned initially, and then the observer was presented with two faces and had to choose which face corresponded to the person whose voice was heard in the first phase. In the FV condition, those were presented in an opposite order. Examples of the structure of trials can also be viewed at http://www.his.atr.co.jp/∼kmiyuki/face-voice.html.

Procedure section). This change also ensured that matching could not be based on the accidental properties of a particular utterance, as might have been the case if we had taken video and audio from the same recording for both phases. Parallel research with the same XAB procedure has shown that the task is possible when the same word or sentence is used for both the first and second phases [13–15]. For example, Lachs [14] has shown that the auditory and visual components for production of the word "*cat*" can be matched when played forward, but not when the stimuli are played backward. Rosenblum and colleagues [15] demonstrate identity matching for different repetitions of the same sentence, but the use of different sentences provides an additional test. In the second series of experiments, we focus specifically on the issue of sentence overlap.

The first experiment constituted a simple test of whether it was possible to match identity across modality with full audio and video information available. As can be seen from Figure 2, performance for the *forward* condition (presentation of unmodified visual and auditory stimuli) was better than chance for both face to voice (FV: 61%; $t_{23} = 3.0$, $P < 0.01$) and voice to face (VF: 62%; $t_{23} = 4.3$, $P < 0.01$) matching. The result shows that, while performance was far from perfect, there is common information that can be used to match identity despite the change in modality. The fact that we used similar but not identical sentences shows that the information used is not entirely utterance or content specific. In this experiment, there was no significant difference in performance between FV and VF conditions ($P > 0.1$). In order to test whether the ability to match identity across modality was dependent on a few particularly distinctive people, we performed an items analysis. Consistent with the subject analysis, performance was above chance for both FV (66.2%; $t_{39} = 4.6$, $P < 0.001$) and VF (62.3%; $t_{39} = 3.4$, $P < 0.01$) trials, suggesting that the effect could be generalized to other items and that it is not just a function of a few particularly distinctive items.

In the second experiment, all auditory and visual stimuli were played backward. This manipulation reverses the overall spatiotemporal patterning while leaving many properties, for example, spatiotemporal ranges and amplitudes, unchanged. Voices played backward are well known to be unintelligible, but speaker identity can still be recognized [16]. This is probably because many cues to voice quality, including spectral patterns relating to the fundamental frequency and formants, are preserved. Similarly, while playing degraded video backward is known to affect motion-based recognition [17], faces can clearly still be recognized from movement-independent image properties. For both modalities, backward play affects the overall, global pattern of change while leaving local properties, ranges, and average values unchanged. Task performance then depends on what

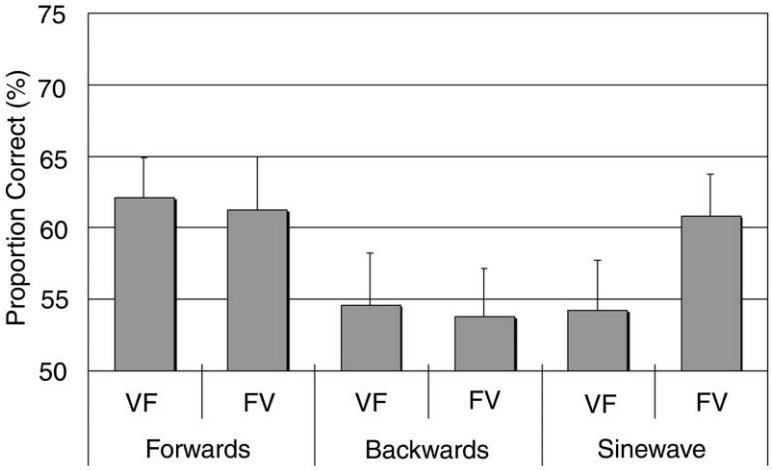

Figure 2. Results of the First Experimental Series, FV/VF Matching of *Forward*, *Backward*, and SWS Conditions

The mean proportion correct is shown as a percentage, with error bars indicating standard errors. The level of chance performance was 50%.

properties are sufficient. In this case, backward play would only be expected to affect performance if the nonlocal auditory and visual spatiotemporal patterns are critical for the task.

As can be seen from Figure 2, direction-dependent properties did appear to be important, in that performance dropped to chance when the stimuli were played backward. One-sample t tests showed that matching performance was at chance for both FV (54%; $t_{23} = 1.1$, $P > 0.1$) and VF (55%; $t_{23} = 1.3$, $P > 0.1$) conditions. Thus, the information used in identity matching must be both time varying and direction specific, and we would expect that no combination of static or direction-independent visual cues with direction-independent auditory properties would be sufficient for this task.

We have run a similar experiment with static images, and, as would be expected on the basis of these results, performance was not above chance: FV (53%: $t_{15} = 0.6$, $P > 0.1$), VF (56%: $t_{15} = 1.5$, $P \geq 0.1$). However, with the use of static images, there is also a drop in the amount of static information available, simply specified in terms of the number of frames, in addition to any effect of the disruption of movement.

In order to further test the importance of *meaningful* spatiotemporal structure, we used *sine wave synthesis* (SWS) [5]. SWS is a resynthesis technique in which broadband spectra are reduced to pure tone sine waves whose frequency and amplitude correspond to the spectral peaks of, usually, the first three regions of spectral prominence (formants) in the original speech acoustics (we generated our sine wave speech automatically from the audio files used in experiment 1 by using code made available by Dan Ellis at http://www.ee.columbia.edu/~dpwe/resources/matlab/sws/. In this code, Linear Predictive Coding was used to estimate the parameters used for sine wave speech. We used sine waves corresponding to the first three harmonics for our resyntheses). Thus, SWS preserves coarse spatiotemporal information but eliminates fundamental frequency and provides few cues to voice quality. Despite the limited and unnatural-sounding information available, SWS is sufficient to convey both meaning and identity in auditory perception tasks [5, 12, 18].

Figure 2 shows that observers were able to match the speaking faces on the video to SWS (FV) at levels significantly better than chance (61%; $t_{23} = 3.7$, $P < 0.005$). This indicates that identity-specific information corresponding to that encoded from the video of the face is available from the spectrally reduced SWS. This suggests the importance of overall spatiotemporal patterning for this task. It is also possible to limit visual information in an analogous way by presenting point-light stimuli rather than full video. In this case, people are able to match a voice to point-light faces [15]; again, this matching ability emphasizes the usefulness of spatiotemporal over static structural information for this task.

In this experiment with SWS, performance in VF trials was not significantly above chance (54%; $t_{23} = 1.2$, $P > 0.1$), suggesting that the information specifying identity was not encoded from sine wave speech. One possible explanation for this failure is that processing resources and attention are primarily devoted to recovering meaning when presented with SWS, and this limits encoding of identity. The results may also reflect a general asymmetry between FV and VF matching, whereby it is more usual to recover information about what is being said from the movement of the face than to recover movement of the face from what is being said.

The three experimental conditions reported thus far — *forward*, *backward*, and *sine wave synthesis* — have highlighted the importance of overall, direction-dependent spatiotemporal structure for this task. In the following experiments, we focused on the extent to which performance is determined by the overlap in what is said between the two phases. While the slightly different sentences used in all the experiments reported so far ensure that the effect is not dependent on an exact match of content, the overlap between the two phases may still be an important factor. Thus, for the following experiments, we used both identical and very different sentences for A/B and X (see the Experimental Procedure section for the sentences used). At one extreme, the sentences used differed in terms of their syllables, words, positions, and total duration (A and B, 2.9 s; X, 3.5 s), while at the other extreme, audio and visual stimuli were taken from the same recording of the same sentences. All faces and voices were played forward with full audio and visual information available.

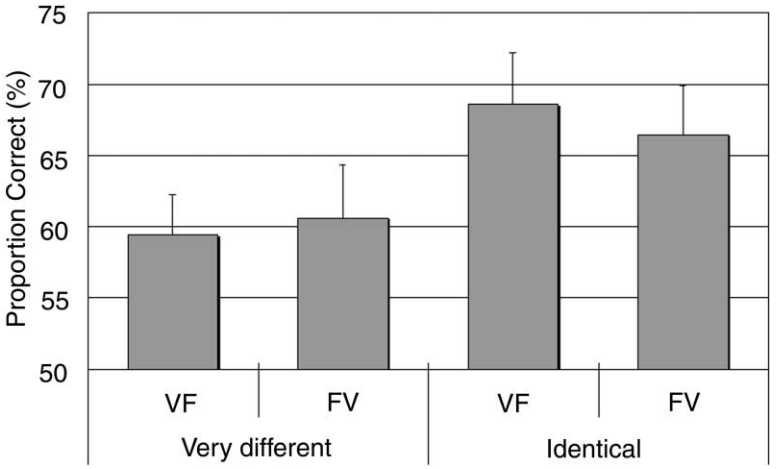

Figure 3. Results of the Second Experimental Series, FV/VF Matching for Very Different Sentences and for Identical Sentences

The mean proportion correct is shown as a percentage, with error bars indicating standard errors. The level of chance performance was 50%.

As can be seen from Figure 3, in the *very different* condition using the *short* and *long* sentences, performance was still above chance for both FV (60%; $t_{17} = 2.64$, $P > 0.1$) and VF (59%; $t_{17} = 2.8$, $P > 0.1$) conditions, and the levels of performance were quite similar to those shown previously for the *forward* condition. This confirms that the ability to match identity is not dependent upon sentence equivalence. Instead, it must reflect more general differences between individual speakers.

In the final experiment, audio and video from the longest sentence available (sentence [c] in the Experimental Procedures section) was used for both X and A/B. One aim of this was to determine the maximum level of performance possible in this task by using the stimuli available. Predictably, performance was above chance for both conditions: FV (66%; $t_{13} = 3.1$, $P < 0.01$) and VF (69%; $t_{13} = 6.0$, $P < 0.0001$). Also, as expected, performance levels were the highest of all the experiments reported. However, a two-way, mixed design ANOVA with sentence similarity and learned modality as factors showed only a marginal main effect of sentence similarity ($F_{1,30} = 3.8$, $P = 0.06$) and no interaction ($F_{1,30} = 0.118$, $P < 0.73$).

Thus, the results show quite stable levels of performance and suggest that the information used is a persistent characteristic of the perception of the speaker and is not limited to a particular utterance or overlap of words. The marginal effect of similarity suggests that utterance-specific cues may convey an additional advantage but that the overlap is not essential. Part of the marginal advantage may have been associated with increased duration anyway: the sentence used in the identical condition was longer than the pairs of similar sentences used (3.9 s as opposed to 2.9 s). However, the results suggest that neither of these factors is critical. Indeed, the levels of performance are similar for both the same repetition of a single syllable word and for different repetitions of the same sentence [13, 15].

In summary, people are able to match the identity of unfamiliar people across auditory and visual modalities at levels that were far from perfect but were significantly above chance. The information used must be either modality independent or translatable between modalities. The information is time varying, direction dependent, and of the coarse spatiotemporal type preserved by sine wave speech.

The facial motion and acoustics associated with speech production are both constrained by structure and time-varying configuration of the vocal tract [4, 19]. The resulting link between auditory and visual speech information allows us to match the identity of faces and voices as well as allows us to speech read. Differences in the way people speak as well as the words that they say are reflected in their voices and in the way their faces move. We are sensitive to these differences and can recover them despite a change in modality. The experiments reported here show that the visual information used for this task, like the auditory information, is inherently time varying. Reversing temporal direction radically alters the dynamics of natural speech, and it is this, as opposed to image-based or other temporally localized properties, that appears to determine performance.

Playing the stimuli backward also clearly affects the perception of speech, and a related issue for future study is to determine the extent that identity matching depends on understanding what is said. While all of the experiments reported to date have involved native speakers and perceivers of either Japanese (this study) or English (other studies), preliminary experiments with Japanese stimuli with English observers suggest that the effect may not be dependent on language. English participants were able to perform better than chance at matching Japanese FV (24 participants, 58%; $t_{23} = 2.9$, $P < 0.01$) stimuli. Thus, results show that identity matching does not depend on perceiving exactly the same words in both modalities or on understanding what has been said. Instead, individual variation in *nonverbal* cues, for example, those influencing prosody and intonation, may be sufficient to specify identity across a change in modality even for an unknown language. Generalization across language would also be consistent with multimodal production work in our laboratory [20] that suggests that the laws linking the movement of the face to the production of sounds are largely determined by the mechanics of sound production and are not specific to particular languages.

Clearly, there are many other possibilities for future

work. It is not just speech and identity that are encoded in both the face and the voice. Other examples include attractiveness, distinctiveness, and the sex of the speaker. Perception of these may also reflect bimodal properties. For example, there is already evidence that perception of emotion in face and voice is linked [21]. While recognition in the independent modalities would make matching of many of these attributes trivial, correlations of ratings or measures of brain activity may be informative. The critical question as to the precise nature of the dynamic information used for all these tasks, including audio-visual speech recognition, also remains. Currently, we are using graphical methods to study the contribution of different types of head, face, jaw, and mouth movement for audio-visual speech perception, and similar techniques, or the presentation of parts of faces, could help to specify the cues used. Lastly, performance, although consistent, was relatively low for this task, and ways of raising the level are clearly desirable. If the critical information can be determined, it may be possible to emphasize or exaggerate it, thereby enhancing performance. Alternatively, increased exposure during both stages, perhaps to a variety of material, would give more scope for individual variation and enhance performance.

To summarize, the auditory and visual information derived from natural speech are closely linked, and this linking affords recovery of speaker-specific information across a change in modality. The generality of the effect across different sentences shows that identity matching does not require repetition of either structure or content, though increased overlap may provide additional cues. The information used is time varying and direction dependent, suggesting that it is closely tied to the dynamics of natural speech production. Thus, this information not only crosses modalities but also bridges the gap between perception and production.

Experimental Procedure

Face and Voice Stimuli

For stimuli, we videotaped the voices and faces of 20 male and 20 female, adult, native Japanese speakers. Each speaker was recorded reading the following three Japanese sentences from a teleprompter: (a) *Oji-san ha(wa) yama-e shibakari-ni ikimashita*, (b) *Oba-san ha(wa) kawa-e sentaku-ni ikimashta*, and (c) *Hoteru tte Netto yoyaku no houga yasuku tomarerundesho?* Sentences (a) and (b) were used as the X and AB sentences in experiment 1 and as the "short" sentence in experiment 2. Sentence (c) was used as the "long" sentence in experiment 2. Faces and voices were digitized separately from the videotape (Betacam SP) into a videodisk array (Accom 2Xtreme). The average sentence lengths were 2.85 s, 2.85 s, and 3.92 s for sentences (a), (b), and (c) respectively. All recording and playback was in NTSC format at 29.97 frames/s.

Participants

A total of 24 observers participated in each condition of *forward*, *backward*, and *sine wave speech* in experiment 1, and 18 observers participated in the *short-long* condition and 14 observers participated in the *long-identical* condition in experiment 2. All observers were aged 18–33 and were native Japanese speakers unfamiliar with the people used as stimuli and had normal or corrected-to-normal audio-visual abilities.

Design

All experiments consisted of 20 trials and followed the basic design illustrated in Figure 1. For each observer, half the speakers were used as targets and half were used as distractors. Face to voice and voice to face trials were blocked separately, and the order of blocks was balanced across observers. If sentence (a) was used as the first phase, then (b) was used as the second phase and vice versa in experiment 1. Sentences in the first and second phases were also balanced across observers.

Procedure

In each trial, observers first saw a face (or a voice), X, followed by two voices (or faces), A and B, presented sequentially. Their task was to choose which of the second pair of stimuli, A or B, was derived from the same person as the initial stimulus, X. In the SWS experiment, we familiarized subjects with SWS prior to testing by presenting SWS stimuli paired with their original voices. These speakers used as examples were not used in the actual experiment. For all experiments, presentation was directly from the Accom videodisk array and was controlled by a SGIO2. Audio stimuli were presented on Sennheiser headphones (HD270), and visual stimuli were presented on a color video monitor (SONY PVM-20M4J). Faces were about 12 cm in width on the screen (filled by blue back) and, viewed from a distance of 80 cm, subtended a visual angle of approximately 8.6°. Audio stimuli were presented monaurally to the right ear in all conditions, except for sine wave speech, which was presented binaurally. Responses were made orally immediately after each trial and were recorded by the experimenter.

Acknowledgments

This research was supported in part by the Telecommunications Advancement Organization of Japan. We would like to thank lab members and three anonymous reviewers for many helpful suggestions.

Received: May 14, 2003
Revised: August 14, 2003
Accepted: August 14, 2003
Published: September 30, 2003

References

1. McGurk, H.a.M. (1976). Hearing lips and seeing voices. Nature *264*, 746–748.
2. Driver, J. (1996). Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading. Nature *381*, 66–68.
3. Munhall, K., and Vatikiotis-Bateson, E. (1998). The moving face during speech communication. In Hearing by Eye, Part 2: The Psychology of Speechreading and audiovisual speech, R. Campbell, B. Dodd, and D. Burnham, eds. (London: Taylor & Francis - Psychology Press), pp. 123–139.
4. Yehia, H.C., Rubin, P.E., and Vatikiotis-Bateson, E. (1998). Quantitative association of vocal-tract and facial behavior. Speech Commun. *26*, 23–44.
5. Remez, R.E., Rubin, P.E., Pisoni, D.B., and Carrell, T.D. (1981). Speech perception without traditional speech cues. Science *212*, 947–949.
6. Dodd, B., and Campbell, R. (1987). Hearing by eye: the psychology of lip-reading (London: Erlbaum).
7. Vatikiotis-Bateson, E., Eigsti, I.M., Yano, S., and Munhall, K.G. (1998). Eye movement of perceivers during audiovisual speech perception. Percept. Psychophys. *60*, 926–940.
8. Vatikiotis-Bateson, E., and Yehia, H. (1996). Physiological modeling of facial motion during speech. Trans. Tech. Comm. Psychol. Physiol. Acoust. *H-96–65*, 1–8.
9. Knight, B., and Johnston, A. (1997). The role of movement in face recognition. Vis. Cogn. *4*, 265–273.
10. Lander, K., Christie, F., and Bruce, V. (1999). The role of movement in the recognition of famous faces. Mem. Cognit. *27*, 974–985.
11. O'Toole, A.J., Roark, D.A., and Abdi, H. (2002). Recognizing moving faces: a psychological and neural synthesis. Trends Cogn. Sci. *6*, 261–266.
12. Remez, R.E., Fellowes, J.M., and Rubin, P.E. (1997). Talker iden-

tification based on phonetic information. J. Exp. Psychol. Hum. Percept. Perform. *23*, 651–666.

13. Lachs, L. (2002). Vocal tract kinematics and crossmodal speech information. In Research on Spoken Language Processing. PhD thesis, Indiana University, Bloomington, Indiana.
14. Lachs, L. (1999). A voice is a face is a voice: cross-modal source identification of indexical information in speech. In Research on Spoken Language Processing (Bloomington, Indiana: Indiana University, Speech Research Laboratory, Department of Psychology), pp. 241–258.
15. Rosenblum, L.D. (2002). The perceptual basis for audiovisual speech integration. In Internation Conference on Spoken Language Processing, J.H.L. Hansen and B. Pellom, eds. (Adelaide, Australia: Causal Productions Pty Ltd.), pp. 1461–1464.
16. Van Lancker, D., Kreiman, J., and Emmorey, K. (1985). Familiar voice recognition: patterns and parameters. Part I. Recognition of backwards voices. J. Phonetics *13*, 19–38.
17. Lander, K., and Bruce, V. (2000). Recognizing famous faces: exploring the benefits of facial motion. Ecol. Psychol. *12*, 259–272.
18. Fellowes, J.M., Remez, R.E., and Rubin, P.E. (1997). Perceiving the sex and identity of a talker without natural vocal timbre. Percept. Psychophys. *59*, 839–849.
19. Vatikiotis-Bateson, E., Munhall, K.G., Hirayama, M., Lee, Y.C., and Terzopoulos, D. (1996). The dynamics of audiovisual behavior in speech. In Speechreading by Humans and Machines, Volume 150, M. Hennecke, ed. (Berlin: Springer-Verlag), pp. 221–232.
20. Yehia, H.C., Kuratate, T., and Vatikiotis-Bateson, E. (2002). Linking facial animation, head motion, and speech acoustics. J. Phonetics *30*, 555–568.
21. de Gelder, B., and Vroomen, J. (2000). The perception of emotions by ear and by eye. Cogn. Emot. *14*, 289–311.